DOCUMENT RESUME

ED 225 577 IR 050 059

AUTHOR Huang, Jack Kai-tung; And Others
TITLE The Design of the CCCII and Its Application Considerations in Library Automation.
PUB DATE Aug 81
NOTE 21p.; Paper presented at the Annual Meeting of the International Federation of Library Associations (Leipzig, East Germany, August 18-23, 1981). Best copy available.
PUB TYPE Reports - Descriptive (141) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Chinese; *Data Processing; Digital Computers; Foreign Countries; Ideography; Library Networks
IDENTIFIERS Character Codes; *Chinese Character Code for Information Interchange; *Information Interchange Codes; Standardization; Taiwan

ABSTRACT

This paper presents the major characteristics of the Chinese Character Code for Information Interchange (CCCII) and indicates its intended application for the interchange of Chinese information among computer systems and communication facilities, especially in library networks. It is considered sufficient for present day library applications, business data processing, and even census application. The selection of CCCII's graphic symbols and its ordering of Chinese characters are discussed, as well as such application considerations as the achievement of coding and processing efficiency, the means of handling variant forms and pronunciations of Chinese characters, and the sorting and searching of Chinese lists. Mention is made of ongoing work on a Chinese cross-reference database with an authority file of Chinese characters and associated intelligent software. Various tables and charts and an 11-item bibliography are provided. (Author/ESR)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

139/UNP/8E

The Design of the CCCII and its Application considerations in Library Automation

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☑ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED225577

By

Jack Kai-tung Huang, Prof., Dept. of C. S. MCC
Ching-chun Hsieh, Prof., Dept. of E. E. NTIT
Chung-tao Chang, Prof., Dept. of E. E. NTIT
Chen-chau Yang, Dir., Software Center, III

This paper presents the major characteristics of the Chinese Character Code for Information Interchange and its application considerations while applied to an Chinese library computer system. The considerations include achieving coding efficiency and processing efficiency, and the ways of handling variant forms of characters, handling various pronunciations of a character, sorting and searching of Chinese lists e.t.c.

IR 050 959

meeting 117

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Carol Henry
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

While selecting a code scheme for a Chinese information processing system, the following problems are unavoidably needed be clarified for system serviceability and efficiency:

1. the code length of characters,
2. how the code can help to search characters by various search keys, such as pronunciation, radical and stroke count e.t.c.,
3. how the code can help to associate variant forms of a character,
4. how the code can help to associate different pronunciations of a character, and
5. how the code can be adapted to various Chinese input sehemes available in the market.

In CCCII, users are recommand to select a subset of CCCII of 2-byte code length for those applications where storage space and processing efficiency are priorities. For considering the problems stated in item 2 to 5 above, the coding scheme of CCCII itself has already designed to solve a part of the stated problems. The above problems can be completely and effectively solved by applying CCCII and an associated cross-reference data-base of the characters in CCCII. The cross-reference data base design for CCCII is an authority file system of characters and some associated intelligent software to handle file operations and some basic data processing functions, such as searching, sorting, matching, and code convertions of characters.

In this paper, some major characteristics of CCCII will be presented first. Then, some application consideration of CCCII, as mentioned above, will be discussed.

1. Introduction

The Chinese Character Code for Information Interchange (abbr. CCCII) is designed for interchange Chinese information among computer systems and communication facilities, especially in library networks. The volume I of CCCII, which contains 4807 most frequently used Chinese characters was published in April 1980. The volume II of CCCII, which contains more than 33 thousands characters, was announced in February 1981.

This set of standard code is basically an extension of the 7-bit code set of ISO 646. Provision of Chinese characters is made by three 7-bit code vectors according to the multi-byte extension specified by ISO 2022. The CCCII is trying to code all Chinese characters, including all variations and simplified forms of each Chinese character, into the three 7-bit-bytes code space. But, there are too many characters for one volume and too much work for one batch. Therefore, we will subsequently publish a series of Chinese and English version of the code standards for different sub-sets of Chinese characters according to the frequency of usage as well as the classification of the forms of each character. This code is the most general arrangement to cover a wide spectrum of applications. For a specific application of limited Chinese characters, one byte of the code may be omitted. It is considered as user's option to use a subset of CCCII. The volume I and II of CCCII form a very powerful tool for Chinese information interchange, it is still a suoset of the CCCII. It is expected to take a couple of years to finish the coding of the rest rarely used characters to make the code completed. That is, make it capable to handle any Chinese information that would be processed by computers.

2. The graphic symbols collected in CCCII

The 4807 characters cited in volume I of CCCII are the most frequently used ones. It is believed that almost all applications need to refer to the set which as published by the Ministry of Education (at Taipei) on May 7th, 1979. The volume II is direct extention of volume I by the number of characters. The total characters collected in those volumes include:

(1) The 4807 most frequently used characters;
(2) 17187 less frequently used ones and
(3) 11008 variant forms corresponding to the characters in items (1) and (2).

Those characters covers all characters being used by the present-day data processing applications in Taiwan. characters for library applications, standard characters announced by the Ministry of Education and variations being used in other Chinese communities outside Taiwan, such as 3285 simplified characters. Table 1 shows from where those characters are collected. It is believed that this collection would be sufficient for present-day library applications, business data processing and even census application.

Other symbols collected in volume II of CCCII include:

1. 214 radicals
2. 42 mandarin pronuncation symbols
3. 53 numbering symbols
4. 35 punctuation symbols
5. 24 business symbols
6. 56 mathmatics symbols
7. 22 other symbols

Table 1: The survey of characters for general data processing

item	Resource (Company Name or Book Title)	Number of characters (Approx.)	Applied fields
1	Taiwan Automation Co.	9,600	These systems have been applied to the data processing works in the following areas: .Electricity .Library .Gas .small business .Telephone company .Banking .Government .Water Organization .Tax .Schools .Police
2	Wang Industrial Co.	10,499	
3	Fincial & Tax Center	11,000	
4	IPX Taiwan Ltd.	9,600	
5	Feng-Chia University	16,000	
6	National police Administration	11,825	
7	Characters for Libraries	15,000	
8	Chiao-Tung University	9,129	Basic character set for all Chinese data processing application and researches
9	The code book of the three corner coding method	8,800	The most popular one and well adapted by users.
10	The Comprehensive Dictionary of Chinese Character Index	9,600	
11	Chinese Characters for Telegraph Code	8,000	for data transmissions
12	Characters for Elementary school	4,600	for teaching
13	The ministry of Education	12,701	These are the standard forms so far published

3. The structure of CCCII

3.1 Ordering of Chinese characters

Within each subset of CCCII, the ordering of characters is arranged according to the Kang-hsi radical sequence first, and then by its stroke-count. For those characters with the same radical and stroke-count, they are sub-arranged by stroke-order. The precedance of strokes is shown by the following descending order:

(1) A dot (丶Tien)
(2) A horizontal stroke (一Heng)
(3) A vertical stroke (丨chih)
(4) A stroke down to the left (丿p'ieh)
(5) A stroke down to the right (乀Na)

By the above arrangement, the CCCII code itself contains a consistant ordering relation for the following search keys: radical, stroke count and stroke sequence. This characteristics is very important and effective while data processing efficiency and serviceability are considered. It also provide a way to obtain a better organization of the software for Chinese data processing systems. The JISc6226 developed by Japan and the code by PRC do not have this consistent ordering characteristics. Therefore, for doing any kind of searching, extra information associated with their code must be provided and extra processing is required. An typical example is to sort two characters, one from the first character group and the other from the second character group. Because the code for the character from the first group contains only the ordering information by its pronunciation, while the character from the second group containsthe ordering information by its radical and stroke count, therefore, we can't ordering than by any key if no other information is suplemented and, apparently, more processing work is required.

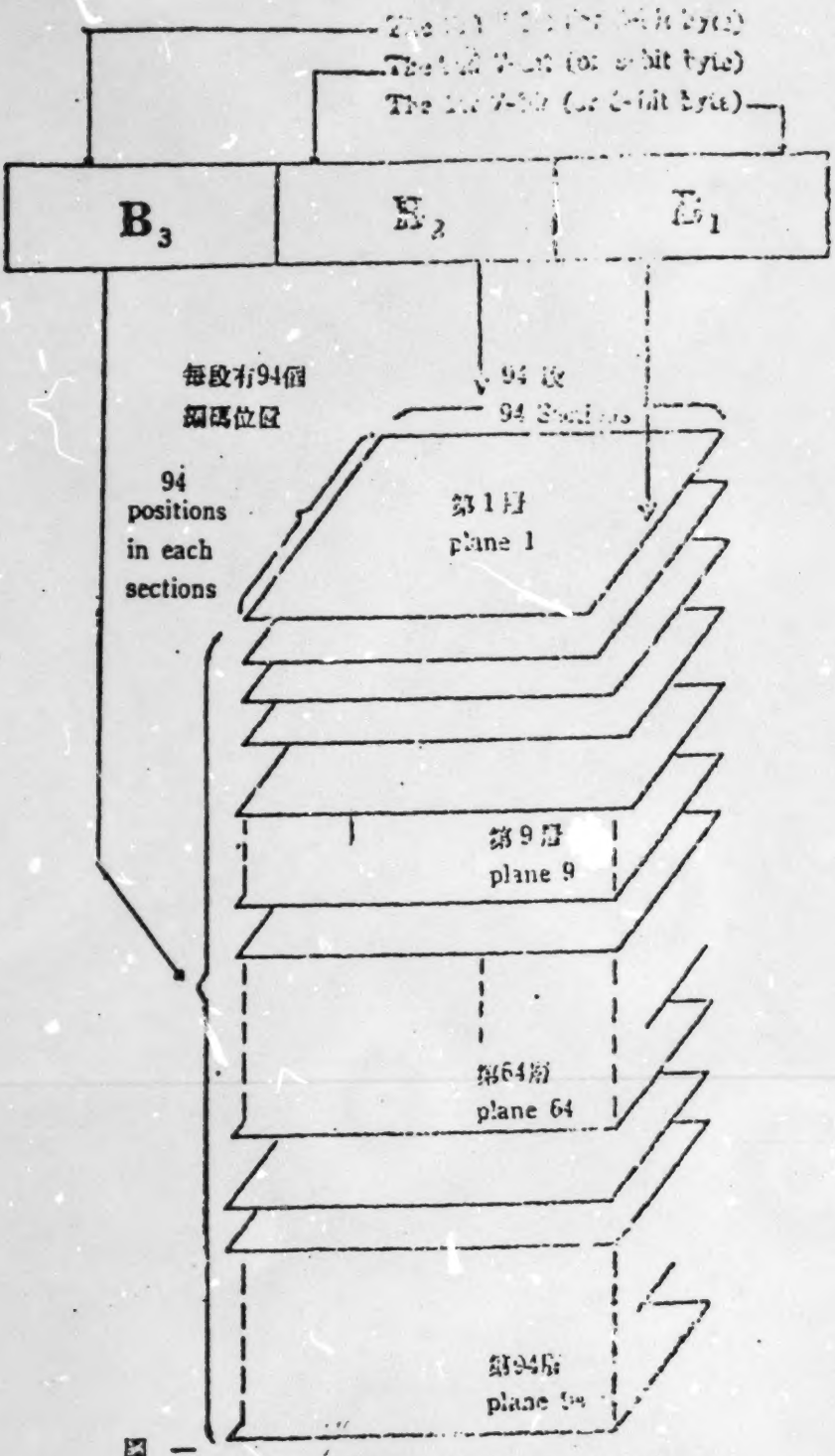
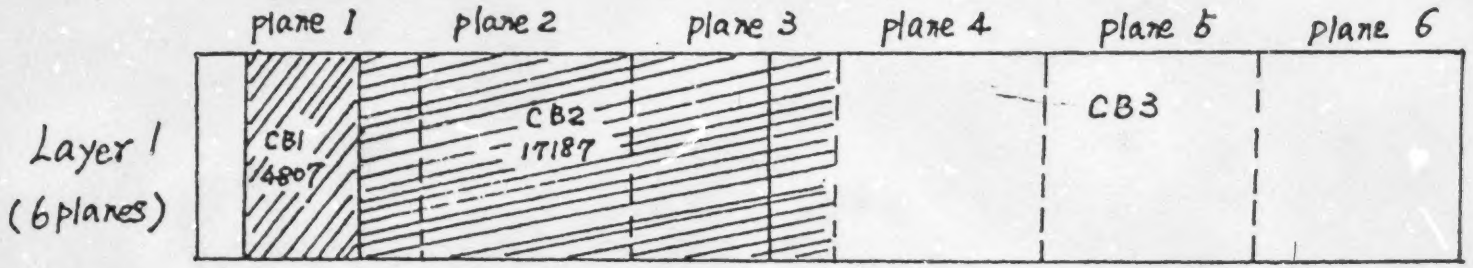


Figure 1 三度空間 94×94×94個編碼位置結構圖
The 3-dimensional structure of the whole 94×94×94 code



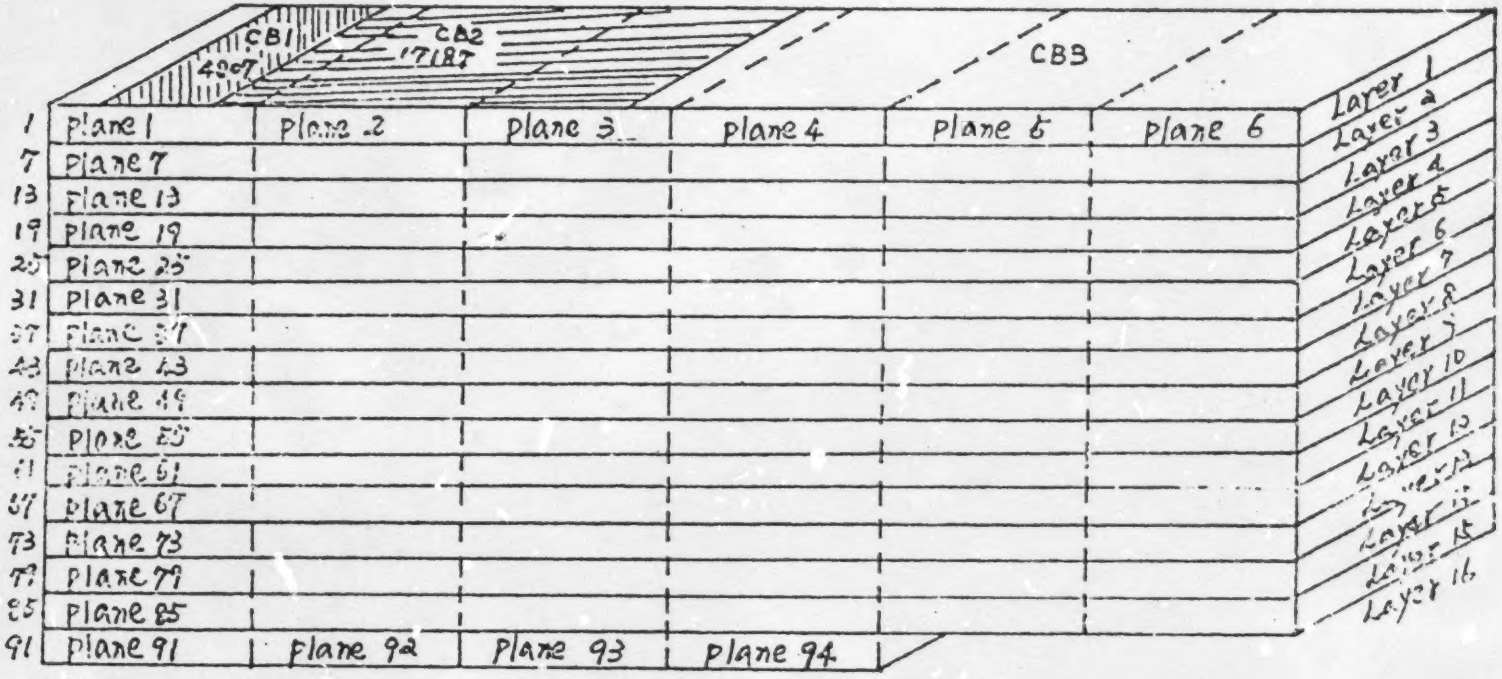


Figure 2: A schmatic diagram of CCCII structure

In CCCII, the codes assigned to the variant forms are desinged to have identical two right-most bytes, B_2B_1, as the code of its corresponding character. In other words, the variant forms are placed at the same section and position just as its corresponding character occupied but in different planes. By this arrangement, variant forms can be uniquely identified and easily be programmed interchangably while needed. An example of variant forms is shown in Table-2.

The codes of variant forms are illustrated in Figure 2. In Figure 2, started from plane 1, for each 6 consecutive planes form a layer. The first layer is the layer for normal characters. From the second layer on are assigned for the variant forms and for future expansion.

4. Selection of 2-byte subset of CCCII

The CCCII is intended to serve as an universal tool for Chinese Information interchange. Therefore, all symbols used by Chinese language and information handling will be collected and coded. But, from user's point of view many applications do not need such a complete set of characters. In order to fulfull the above two to controversial requirements, the CCCII is so designed that the user can select a proper subset of CCCII to make the code more suitable for his own need.

The subsets of CCCII with 2-byte subcode length are listed in Table 3. Under this circumstances, by agreement between the interchanging parties, only 2-byte is needed to a represent a Chinese character. Please note that the code structure of all those subsets of CCCII is fully ISO 646 compatible.

5. CCCII used as internal code

The CCCII can be used directly as the internal code for processing and storing Chinese information. For applications of high volume of Chinese information, the radix-94 code described in this section can compress storage space considerably.

As shown in Figure 2, the first 6 planes in the first layer have totally 94x94x6=53016 positions. Since this number is less than 2^{16}=65536, the 3 byte code in the first layer can be compressed into a 16-bit or two 3-bit bytes as an unsigned binary number by applying radix-94 conversion. This conversion is a one to one, two way and unique conversion. Therefore, this code can be used as internal code for mass data storage to save storage space if the characters used are limited within 6 planes.

The transformation rules are as follows

1. From CCCII to radix-94 code

   $(B_3-33)x94^2+(B_2-33)x94+(B_1-33)$ —— Equivalent radix-94 code

2. From radix-94 code to CCCI

   $\frac{\text{Radix-94 code}}{94} = A_1$, with remainder=$B_1$,

   $\frac{A_1}{94} = A_2$, with remainder=B_2, and A_2=B_3.

6. Code-length compression by commands

The first sections of each plane are reserved for commands used by Chinese data processing or defined by user. For some applications that the full CCCII must be used, user may define commands to switch among planes and leave the Chinese character code be two-byte. By this arrangement, the length of the code for a block of characters may be significantly reduced.

異體字形 CCCII 編碼結構實例

	6 0	6 0	6 0	6 0	6 0	6 0	6 0	6 0	6 0	6 0
	4 9	4 B	4 C	4 E	4 D	4 F	5 0	5 1	5 2	5 4
2 1	頖	頤	顐	額	顏	題	顑	顓	顠	顡
2 7	頖	頤	顐	額	顏	題	顑	顓	顟	顣
2 D	頖	臣		額	顏	題	顫		顠	顡
3 3	頖				顏		顎			
3 9	頖				題		顎			
3 F	頁						顱			

3rd byte
B_3

B_2
2ed Byte

B_1
1st Byte

normal form
此列為通用體

simplified form
此列為大陸簡體

other variations
此下四列為同義異體

Table 2: An example of the table of variant forms and their associated CCCII codes.

Table 3: The subset of CCCII with 2-Byte subcode length

Subset	No. of characters	Character set Name	Publication	Sub-code length	Compatibility	User defined positions	Typical Applications
CCCII-1	4807	The most frequently used characters (plane 1)	Vol. 1 of CCCII	Two 7-bit Bytes	Fully ISO646 Compatible	653	Teaching, writing, newspaper printing, word processing
CCCII-2	22000	The character set for general data processing (plane 1,2 & 3)	Vol. 2 of CCCII	two 8-bit Byte	Fully ISO646 Compatible	658	Business Data Preprocessing Library, Census applications
CCCII-3	≈33600	Complete first 4 planes	Not yet	two 8-bit Byte	Fully ISO646 Compatible	658	For more specific applications

As an example, let define two switch commands as follows:

Command name	Paramters	Comment
Switch plane, locked	Plane number	4 bytes more for each switch or a group of character
Switch plane, unlocked	Plane number	4 bytes more for each switch or a simple character code.

Then the third byte B_3, the pointer of the plane number, may be ommitted for all the characters. Assume the command name and parameters are 2-byte each, and there are 95% of chance to use those characters in plane 1. Then the weighted average of code length can be computed as follows if only the unlocked switch plane command is used:

average code length = 2 x 95% = 6 x (1 -95%)
= 1.9 = 0.3
= 2.2 bytes

The 7-bit bytes used to code command names and parameters are all belong to the set G0 defined in ISO 646. This makes the switching operation fully located within the processing operation of Chinese characters and hence fully ISO 646 compatible.

7. The cross-reference data-base

The Chinese character indexing problem is defined as the problem of identifying, finding or addressing a Chinese character by its proprietary information, such as its pronunciation stroke count, radical telegraph code etc. It is quite clear that the experience of processing alphabets can not be applied here directly. When we are searching for a character, we usually search it by some information associated with the character. Therefore, it is not a simple code match problem, but an associative search problem. As a consequence, we must store and organize all the necessary information of character to support the capability of doing associative search.

主索引檔
Main Index File

CCCII 字形主鏈 image main pointer

部首筆畫筆順 radical, stroke count and stroke sequence

字根形碼 component expression

三角編號 3-corner code

音標鏈 phonetic pointer

異體鏈 varlant form pointer

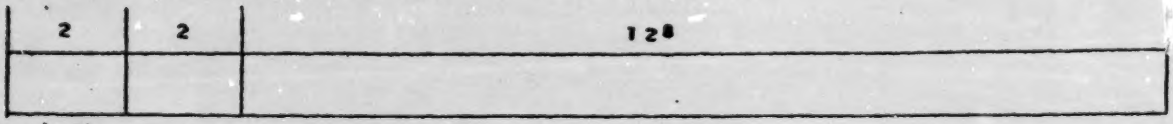
交互鏈 link pointer

其它鏈 other pointers

47 Bytes/Record

235 M Bytes/50000 characters

字形主檔
Stroke Image Main File



主鏈 main pointer

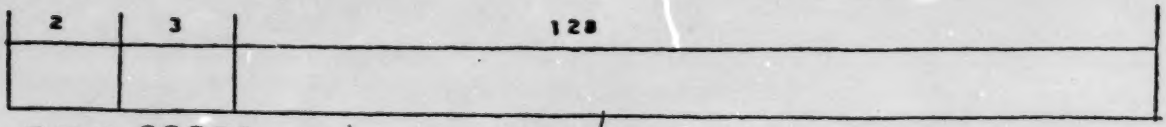
異體鏈 variant forms pointer

字形（32×32點陣） dot matrix image of 32x32

132 Byte/Record

6.6 M Bytes/50000 ch.

異體字形檔
Stroke Image File For Varinat Forms



異體鏈 CCCII pointer

字形（32×32點陣） Dot matrix image

133 Byte/Record

2 M Bytes/15000 Ch.

7.2 Basic Major Functions of the cross-reference data-base

Listed below are the basic functions provided by the intelligent software of the cross-reference data-base. This part of the work has not been completed yet. It is expected to be available by the third quarter of 1982.

1. search a Chinese character and its associated information by using any one of the search key, listed on page 14,
2. Code convertion, among the items listed on page 14, of characters
3. generation of inverted files by any one key listed on page 14,
4. sorting a list of Chinese information according to any one key listed on page 14,
5. selects the character image among nomal form and variant forms,
6. auto-transformation of codes, as an option, among normal form and variant forms while retrieving or searching information from user's text files,
7. accommodation of various Chinese input methods,
8. file maintenance and updating functions, such as add a new field to venders code.

8. Concluding Remark and Acknowledgement

The work reported in this paper is done by the Chinese Character Analysis Group (abbr. CCAG) which is supported by two private non-profit foundations, namely and . It is their support that make this research possible. All the accomplishments so far made are just a partial result of a 5-year research and development project. Detail technique reports are available upon request, and you are welcome to ask to put your name on the mailing list of CCCII.

General References

1. The Chinese Character Analysis Group, CCCII vol I and vol II, NTIT, April 1980 and February 1981 respectively.
2. Ching-chun Hsich et., the Design and application of the Chinese Character Code for Information Interchange International workshop on Chinese Library Automation, NT.T, Taipei, 1981
3. C.T. Chang et., The Design of a Cross-reference Data Base for Chinese character Indexing, International workshop on Chinese Library Automation, NTIT, Taipei, February 1981
4. D.J. Liu, A Comparative Study of Romanization Systems; Internation Workshop on Chinese Library Automation, NTIT, Taipei, February 1981.
5. Jack K.T. Huang et, The Digitized Chinese Dictionary, Taipei, System publication, 1979.
6. Dah-jen Liu et, The comprehensive dictionary of Chinese character index, New York, Asian Associates, 1979.
7. Japan Industrial Standard JISc6226
8. Communist China, Code of Chinese Graphic Character Set for Information Interchange, primary set, 6B2312-80, Beijing, 1981
9. C.K. Plan, A Survey of Various Forms of Chinese Characters, Internation Workshop on Chinese Library Automation, NTIT, Taipei, February 1981.
10. C.F. Chou, Discussion on the Arrangement of Characters Used in Computers from the Viewpoint of Chinese Character Structure and Evolutionary Changes, Internation Workshop on Chinese Library Automation, NTIT, Taipei, February 1981.
11. PSO publications: ISO 646, ISO 2022 and ISO 2375.